Superposition of Molecules: Electron Density Fitting by Application of Fourier Transforms

J. W. M. NISSINK,[1]* M. L. VERDONK,[1] J. KROON,[1] T. MIETZNER,[2] and G. KLEBE[2]

[1]Utrecht University, Bijvoet Center for Biomolecular Research, Padualaan 8, 3584 CH Utrecht, The Netherlands; and [2]BASF AG, Main Laboratory, Carl-Bosch Strasse, D-67056 Ludwigshafen, Federal Republic of Germany

Received 22 March 1996; accepted 22 July 1996

ABSTRACT

In this article a new method is described to superimpose molecules using a crystallographic Fourier transform approach. Superimposed molecules, among other purposes, serve as a basis for three-dimensional (3D) QSAR analyses in drug design and therefore an objective and reproducible method of molecule alignment is of major importance. Fourier data are generated for hypothetical crystals of cubic symmetry for the compounds under consideration. A Patterson-density-based similarity index is used to optimize rotational alignment of the molecules. After optimization of rotational orientation, an electron density derived similarity index is used to further optimize overlap of electron density as a function of translation of the molecules. Both similarity indices are maximized by a simple optimization routine, thus enabling automated superposition. The use of Fourier space offers several advantages. First, rotational and translational parameters can be optimized separately, thus providing a small parameter space. Second, a limited number of data already provide an adequate, continuous description of the electron (or Patterson) density distribution. Third, crystallography provides simple methods to calculate the Fourier transforms that are needed. The resolution of the Patterson (electron)

* Author to whom all correspondence should be addressed at Utrecht University, Sorbonnelaan 16, 3584 Utrecht, The Netherlands.

Journal of Computational Chemistry, Vol. 18, No. 5, 638–645 (1997)

CCC 0192-8651 / 97 / 050638-08

density representation used for superposition can be varied in a straightforward manner. Results are shown for the superposition of two antiviral agents, *2rs1* and *2r04*; the dihydrofolate reductase ligands, methotrexate and dihydrofolate; and a set of three ε-thrombin inhibitors.

Introduction

The assessment of molecular similarity has been acknowledged widely as an important tool in drug design, or, more generally, in elucidating and understanding ligand–receptor interactions.[1–3] Methods that are applied if both the ligand and receptor are known are called *direct design* procedures; molecular complementarity rather than similarity is assessed. Typical methods are docking techniques (ligand–receptor fit analysis) and database searching[4–6] to find three-dimensional (3D) structures that identify templates.[7,8] The similarity-based *indirect design* approach or *active analogue approach*[1] comprises the modeling of small sets of molecules. This can be done by interactive (visual) superposition of molecules, or, in a more objective way, by least squares fitting of geometry,[9] geometry evaluation,[10] or by fitting of molecular shape, molecular surface, electron density, or molecular electrostatic potential.[11–13] Furthermore, linear notation systems have been devised to facilitate molecular comparison.[14,15]

Most approaches work fairly well for a congeneric series of compounds having a common molecular skeleton. Particularly if geometry-based methods are used, difficulties are encountered for compounds with strongly deviating molecular frameworks.

The assumption that a compound fits a receptor cavity leads to the requirement of similarity in *molecular volume*. Molecular volume can easily be evaluated by centering spheres of van der Waals radii at the atomic positions and consecutively evaluating the overlap of active compounds among each other. More elaborate models of volume boundary definition are, among others, solvent-accessible surfaces, as proposed by Connolly.[16] Similarity will also express itself through a similarity in *electron density distribution*[17–19] and through the derived properties of *molecular electrostatic potential* and *molecular electrostatic field*.[1,20] Of these three, the electron density distribution inherently incorporates steric factors.

For two molecules superimposed in space, goodness-of-fit can be assessed by calculation of a similarity index based on electron density. A similarity index, R_C, as proposed by Carbó,[21] is given for two molecules, A and B, with electron density distributions, ρ_A and ρ_B, in eq. (1):

$$R_C^{AB} = \frac{\int_V \rho_A \rho_B \, d\nu}{(\int_V \rho_A^2 \, d\nu)^{1/2} (\int_V \rho_B^2 \, d\nu)^{1/2}} \tag{1}$$

Integration is performed over a certain volume, V. The normalized index ranges from 0 to 1, the latter indicating identity.

An alternative index was proposed by Hodgkin[20]:

$$R_H^{AB} = \frac{2\int_V \rho_A \rho_B \, d\nu}{\int_V \rho_A^2 \, d\nu + \int_V \rho_B^2 \, d\nu} \tag{2}$$

This index better reflects numerical differences in magnitude, whereas eq. (1) only responds well to shape differences (see Hodgkin et al.[20]). Both indices (1) and (2) are normalized squared difference indices.

In this article we present a Fourier-based superposition algorithm based on the overlay of Patterson and electron densities as used in X-ray crystallography. An alternative Fourier-space approach is described by Cooper and Allan and is based on momentum space concepts to avoid difficulties associated with usual position space approaches.[22]

Methodology

X-ray crystallography provides a tool to describe the continuous electron density distribution within a molecule using a limited set of data. For this purpose, the molecule is placed at the center of a cubic unit cell of side a which is part of a hypothetical cubic lattice. The continuous electron distribution is described by a set of reciprocal lattice vectors identified by Laue indices, $\mathbf{h}$ (h_1, h_2, h_3 integers), and their corresonding structure factors, $F_{\mathbf{h}}$. These structure factors provide an image of the real space electron distribution in

Fourier (reciprocal) space according to:

$$F_{\mathbf{h}} = \int_{V_{cell}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}) V \, d\mathbf{r}$$

$$\mathbf{h} = (h_1 a^*, h_2 a^*, 'h_3 a^*) \quad \text{with } a^* = 1/a \qquad (3)$$

Vector $\mathbf{r}$ is the positional vector and V is the volume of the unit cell. The structure factor can be regarded as a complex vector of length $|F_{\mathbf{h}}|$ and phase $\varphi_{\mathbf{h}}$ (*vide infra*). A much faster method to calculate $F_{\mathbf{h}}$ is to sum the contributions of individual atoms in Fourier space. These atomic contributions are known as the scattering factors, f, and they are the Fourier transforms of atomic electron density in real space. Eq. (3) can now be written as:

$$F_{\mathbf{h}} = \sum_j f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j) = |F_{\mathbf{h}}| V \exp(i\varphi_{\mathbf{h}}) \qquad (4)$$

The summation is performed over all scattering centers j (i.e., atoms) at positions $\mathbf{r}_j$. The scattering factors are a function of the lattice vector, $\mathbf{h}$. In X-ray crystallography, the scattering factor of an atom is usually calculated by:

$$f = \sum_{m=1}^{4} A_m \exp\left(-B_m |\mathbf{h}|^2 / a\right) + C \qquad (5)$$

Parameters A, B, and C are derived from high quality quantum-chemical data[23] and were taken from SHELX(76).[24] Substituting the scattering factor simply by the number of electrons of each individual atom would yield a point charge description of the molecule. The resolution $(1/|\mathbf{h}|_{\max})$ of the electron density distribution can easily be adjusted by altering the number of reciprocal lattice vectors taken into consideration.

Similarity Index Calculation

Similarity index calculation was performed by placing the molecules to be evaluated in a hypothetical cubic unit cell that embeds both molecules with a sufficient margin. The structure factors $F_{\mathbf{h}}$, can be calculated easily by summation over all atom contributions according to eq. (4). A similarity index can now be calculated according to:

$$S_{\text{Patt}}^{AB} = \frac{2\Sigma_{\mathbf{h}} |F_{\mathbf{h}}^A|^2 |F_{\mathbf{h}}^B|^2}{\Sigma_{\mathbf{h}} |F_{\mathbf{h}}^A|^4 + \Sigma_{\mathbf{h}} |F_{\mathbf{h}}^B|^4} \qquad (6)$$

In accordance with Friedel's law ($|F_{\mathbf{h}}| = |F_{\bar{\mathbf{h}}}|$), only half of the reciprocal lattice has to be used. The use of $|F_{\mathbf{h}}|^2$ implies a Paterson density distribution, as these are the Fourier coefficients for a Patterson synthesis. The advantages of the application of a Patterson density rather than an electron density is discussed next.

Automated Superposition: Rotational and Translational Alignment

The index, S_{Patt}^{AB}, is calculated according to the protocol given. To optimize overlays, the similarity index is to be maximized automatically as a function of spatial parameters. A most advantageous side effect of the application of the Patterson density is that the similarity index is determined only as a function of the rotational parameters. An optimal rotational orientation is determined by application of a simple optimization routine.

There are several ways to derive the optimal translation vector. As translation of a molecule in the hypothetical unit cell has no effect on the calculation of S_{Patt}^{AB}, the translation may be made ''effective'' by supplying a reference. This reference can be an additional point charge or atom that has a fixed position in the unit cell. Another possibility is to change the space group of the hypothetical lattice to $P\bar{1}$.

Here, we will use a criterion derived from *electron* density overlap. It can be shown that overlap of electron density is maximized by optimizing the criterion:

$$S_{\text{el}}^{AB} \propto \sum_{\mathbf{h}} |F_{\mathbf{h}}^A| |F_{\mathbf{h}}^B| \cos(\varphi_{\mathbf{h}}^A - \varphi_{\mathbf{h}}^B - 2\pi \mathbf{h} \cdot \mathbf{y}) \qquad (7)$$

with φ^A and φ^B phases of structure factors F^A and F^B. Here, $\mathbf{y}$ is the translation vector to be optimized, given the chosen orientations of A and B.

Experimental

The approach just described was implemented in a computer program (QUASIMODI, FORTRAN code available upon request). The performance of the algorithm was tested for five sets of molecules. Rotation was optimized according to the Patterson index [eq. (6)] using a simplex optimization algorithm.[25] For each rotational optimum found, the translation was optimized using the Fourier least-squares residual [eq. (7)]. Calculations were performed on a Silicon Graphics Challenge computer using one 75 MHz IP21 processor.

Final orientations for a pair of molecules were compared to superpositions as derived from crystal structures of the corresponding protein–ligand complexes retrieved from the Brookhaven Protein Data Bank.[26] These ''experimental'' alignments were obtained by superimposing the protein backbones of the different complexes. To quantify the comparison, root-mean-square (RMS) deviations for coordinates of optimized orientations of the second molecule with respect to the experimental superposition are given. All pairs of molecules were optimized from 12 different starting orientations covering the rotational parameter space.

Results

2RS1 AND 2R04

The superposition of two antiviral agents, *2r04* and *2rs1*,[27] taken from the structures of their complexes with one protein of the capsid of the human rhinovirus, was used to check the influence of the resolution on the results. Unit cell sizes are 23 $\pm$ 1 Å for all optimizations. Resolution was varied by varying the size of the reciprocal lattice vector set.

As can be seen in Figure 1*a* and *b*, the overlap function is not altered substantially upon increase of resolution. One of the effects of larger lattice vector sets is smoothing out the artifacts that are introduced at low resolutions.

The RMS value results and computation times are given in Table I. Table I clearly shows that optimizations lead to a small set of optimal overlays that for all resolutions are similar. At every resolution, a distinct maximum similarity is found, in this case near the experimental superposition (as indicated by a low RMS value). Solutions are ranked according to S_{el}^{AB}, reflecting electron density overlap. As can be seen, low resolutions already yield good results; computing times are short. Generally, the number of different solutions is small; equivalent solutions are found more than once, starting the optimization procedure from different initial orientations.

Four best overlay solutions are shown in Figure 2 together with the experimentally derived superposition from the protein–ligand complexes (resolution 4 Å). High RMS values (larger than 12 Å) indicate orientations in which the superposition is reverse. An RMS value of 3.43 Å corresponds to an orientation in which the optimized orientation of the rod-shaped molecule is rotated by approximately 90° about its long axis with respect to the experimental superposition. The 0.63-Å RMS solution is near the protein–ligand alignment as derived from the crystal structures. On visual inspection, the other solutions appear to be adequate and cannot be rejected without further circumstantial evidence.

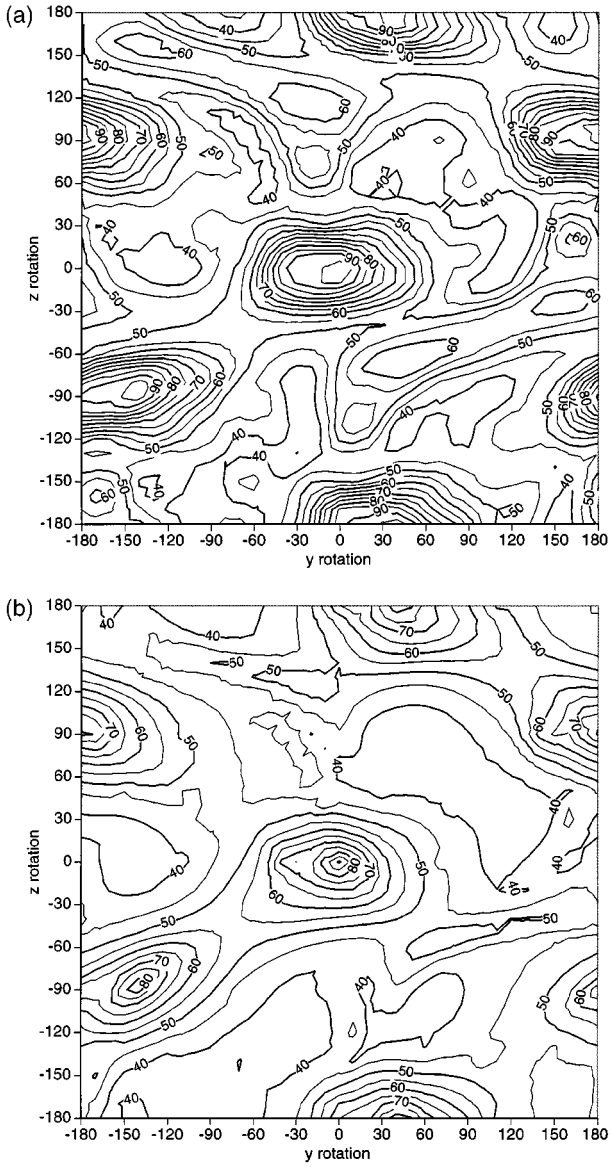

FIGURE 1. Similarity index S_{Patt} contour plots for superposition of *2rs1* and *2r04* with a Laue set of 16 (a) and a Laue set of 3576 (b) (corresponding to resolutions of 11 Å and 2 Å, respectively). The molecule is approximately aligned with the *x*-axis. The *x*-axis rotation is held fixed at 0° (with respect to the protein–ligand crystal structure overlay); similarity index values are given in percent. Contour line intervals are equal for the two plots. The (0, 0) position corresponds to the protein–ligand crystal structure overlay.

TABLE I.

Root Mean Square (Å) Values for Solutions Optimized from 12 Different Starting Orientations of Compounds *2r04* and *2rs1*.[a]

Resolution (Å)	Time (seconds)		RMS values for solutions of overlays *2rs1* and *2rs4* ranked according to similarity index S_{el}						
			1st	2nd	3rd	4th	5th	6th	7th
8	5.4	RMS	0.67	1.30	12.85	3.24	12.88	12.84	
		S_{el}	1.01	1.00	0.97	0.89	0.60	0.51	
6	12.0	RMS	0.58	12.85	13.02	3.29	13.17	13.46	15.01
		S_{el}	1.00	0.95	0.83	0.79	0.55	0.47	0.02
4	45.1	RMS	0.63	12.98	12.84	3.43	13.41	13.90	
		S_{el}	0.95	0.84	0.83	0.67	0.47	0.41[b]	
3	150.4	RMS	0.62	12.99	12.86	12.96	3.32	14.56	14.36
		S_{el}	0.87	0.79	0.59	0.54	0.53	0.24	0.19[b]

[a] Solutions are ranked according to the similarity index S_{el}, and for each solution an RMS value of coordinates (with respect to the protein–ligand superposition) is given. Similar solutions are grouped together. Some solutions were found more than once starting from different initial orientations. Computing times are mean values for one overlay optimization.
[b] Lowest-ranked solution (no overlap) has been omitted.

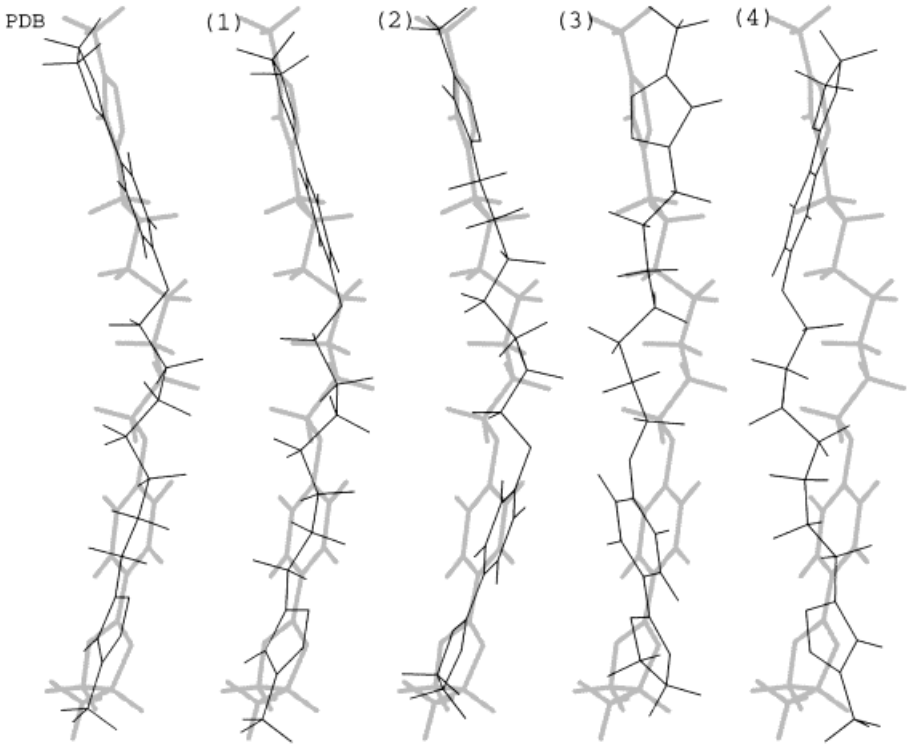

FIGURE 2. Optimized orientations for superposition of *2r04* on *2rs1*. The leftmost orientation shows the crystal structure overlay (''PDB''); nos. (1) to (4) indicate the four best overlays with RMS values of 0.63, 0.62, 12.84, and 3.43 Å, respectively. The template molecule is shown in gray (*2rs1*).

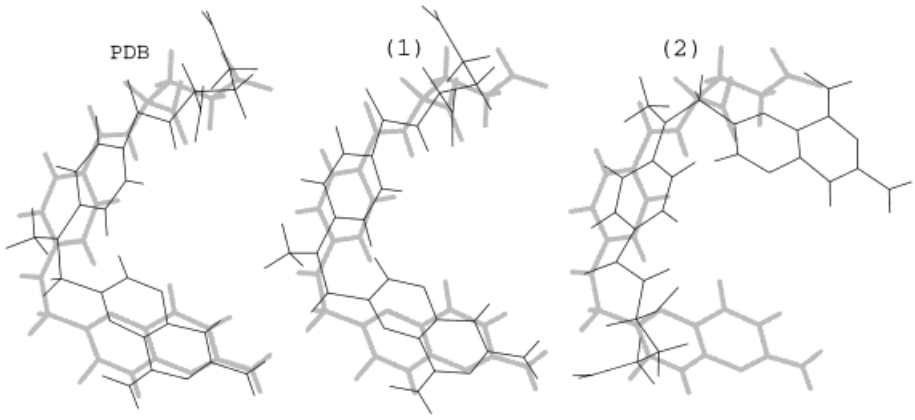

FIGURE 3. Optimized orientations of methotrexate superposed on dihydrofolate. Dihydrofolate is shown in gray. Only the two best superpositions are shown, the experimental alignment is marked ''PDB''. Optimizations were calculated at a resolution of 3 Å.

DIHYDROFOLATE AND METHOTREXATE

As a second case, the superposition of dihydrofolate and methotrexate[28,29] from their complexes with dihydrofolate reductase was used. From the RMS deviation between the optimized set and the crystal structure data it was seen that the orientation of dihydrofolate on methotrexate is only slightly altered for solutions calculated at both high and low resolution. Two major orientations found for the superposition are shown in Figure 3. The remaining alignments correspond to poor overlaps of the electron density, as indicated by the S_{el} similarity index (Table II). The suggested solutions appear very reasonable and the best solution found is consistent with experimental data. It was observed that, at 2 Å resolution, superpositions lead to results with better partial atomic overlap, whereas lower resolution superpositions yield a better overlap of general shape of the molecule.

ε-THROMBIN INHIBITORS

A set of three ε-thrombin inhibitors, MQPA, NAPAP, and 4-TAPAP,[30] was processed at a resolution of 2.5 Å and first-ranked overlay solutions are displayed in Figure 4 for the total set of three. The alignments of MQPA onto 4-TAPAP on the one hand and NAPAP onto MQPA on the other hand perform well as indicated by low RMS val-

TABLE II.

Root Mean Square (Å) Values for Superpositions Optimized from 12 Different Starting Orientations of Compounds Dihydrofolate and Methotrexate.[a]

Resolution (Å)		RMS values for solutions of overlays dihydrofolate and methotrexate ranked according to similarity index S_{el}						
		1st	2nd	3rd	4th	5th	6th	7th
3	RMS	1.28	1.18	8.82	9.56	7.80	5.17	
	S_{el}	0.87	0.86	0.34	0.27	0.24	0.20	
2	RMS	0.87	8.92	6.94	10.02	9.14	3.59	8.81
	S_{el}	0.69	0.24	0.24	0.20	0.20	0.19	0.17[b]

[a] Solutions are ranked according to the similarity index S_{el}, and for each solution an RMS value of coordinates (with respect to the experimental superposition) is given. Similar solutions are grouped together. Some solutions were found more than once starting from different initial orientations.

[b] Three worst solutions ($S_{el} < 0.16$) have been omitted.

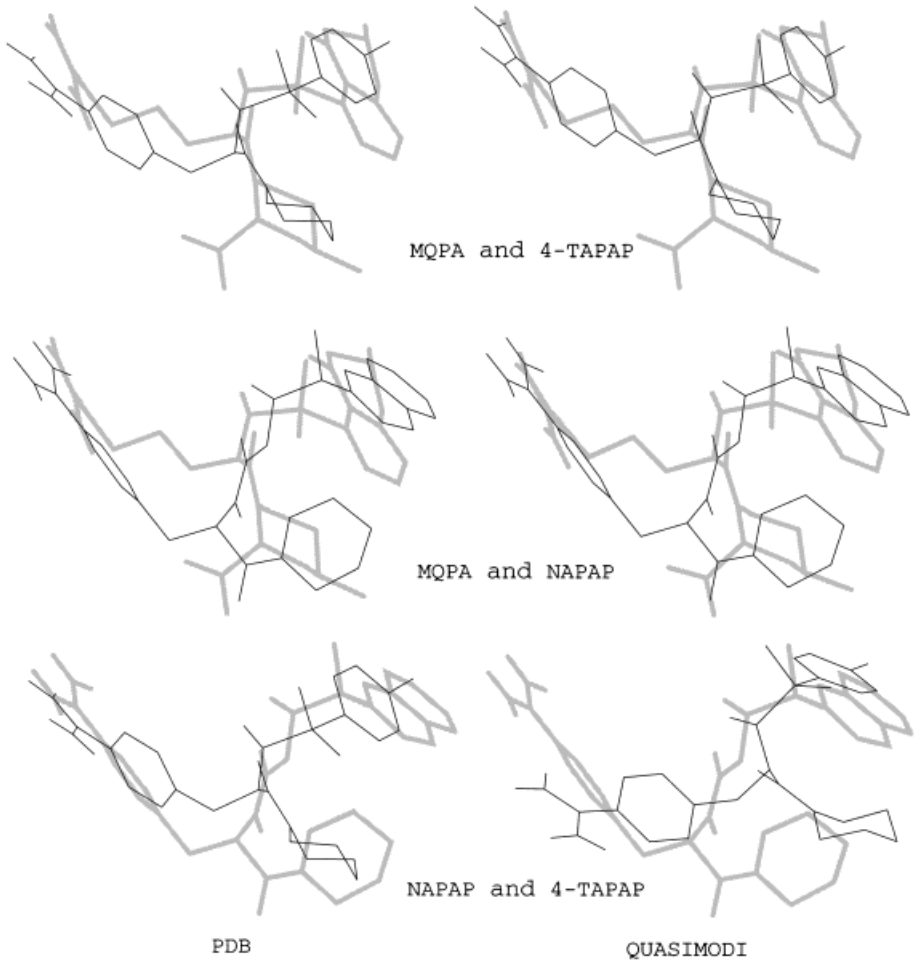

FIGURE 4. Three best superpositions for the set of three thrombin inhibitors. PDB-derived alignments are shown on the left side; computed overlays are shown on the right. The template molecule is shown in gray. The optimizations shown were calculated at a resolution of 3 Å.

ues and relatively high overlap indices (MQPA/4-TAPAP RMS = 0.9 Å, S_{el} = 0.74; *NAPAP/MQPA RMS* = 0.8 *Å*, S_{el} = 0.58). The optimization of NAPAP and 4-TAPAP reveals a higher RMS value (RMS = 2.4 Å) and a lower index (S_{el} = 0.44) than the other two optimizations. However, the overall orientation of both molecules is similar to the overlay deduced from the protein–ligand complexes. In all three optimizations, the best S_{el}-ranked superposition corresponds to the orientation that is observed in ligand–protein structures.

Conclusion

In this article, we proposed a new method to superimpose molecules according to a combination of Patterson- and electron-density-derived similarity indices. An important feature is the reduction of parameters to be optimized simultaneously, as the method implicitly leads to a separation of rotational and translational parameters. Most grid-based approaches suffer from the problem that translation and rotation parameters have to be optimized concurrently in an iterative way, often resulting in a parameter space containing many local maxima.

Apart from its relatively fast calculation (when compared to grid-based methods), another advantage of the proposed algorithm is that it does not require any additional parameterization before data handling, although the choice of resolution needs to be considered. The handling of electron density models in Fourier space is widely applied in crystallography. Although the description of the

electron density in reciprocal space is noncontinuous, one should be aware that the Fourier representation describes a continuous electron distribution in space to a certain resolution. The resolution is determined by the size of the reciprocal lattice vector set used.

The reciprocal space model is very flexible toward modified information. As all information from the reciprocal space is linked to atomic positions, it is fairly easy to introduce a weighting scheme for the atomic contributions. Even fitting of molecular fragments instead of total molecules is easily applied. This would allow a fit focusing on a particular pharmacophore in the molecules, which would further limit computing times.

Acknowledgment

One of us (J. W. M. N.) thanks BASF AG for providing the facilities to work on the model described in this article during a stay at BASF Ludwigshafen during September–October 1994.

References

1. G. R. Marshall, C. D. Barry, H. E. Bosshard, R. A. Dammkoehler, and D. A. Dunn, In *Computer-Assisted Drug Design*, *Vol. 112*, E. C. Olson and R. E. Christoffersen, Eds., American Chemical Society, Washington, DC, 1979, p. 205.
2. C. W. Thornber, *Chem. Soc. Rev.*, **7**, 563 (1979).
3. J. S. Dixon, *Trends Biotechnol.*, **10**, 357 (1992).
4. S. K. Kearsley and G. M. Smith, *Tetrahed. Comput. Meth.*, **3**, 615 (1990).
5. V. J. v. Geerestein, N. C. Perry, P. D. J. Grootenhuis, and C. A. Haasnoot, *Tetrahed. Comput. Meth.*, **3**, 595 (1990).
6. G. Klebe, T. Mietzner, and F. Weber, *J. Comput.-Aided Mol. Design*, **8**, 751 (1994).
7. Y. C. Martin, *Meth. Enzymol.*, **203**, 587 (1991).
8. N. C. Cohen, J. M. Blaney, C. Humblet, P. Gund, and D. C. Barry, *J. Med. Chem.*, **33**, 883 (1990).
9. P. K. Redington, *Comput. Chem.*, **16**, 217 (1992).
10. G. M. Crippen, *Mol. Pharmacol.*, **22**, 11 (1982).
11. A. J. Hopfinger, *J. Am. Chem. Soc.*, **102**, 7196 (1980).
12. M. Marsili, P. Floersheim, and A. S. Dreiding, *Comput. Chem.*, **7**, 175 (1983).
13. C. Humblett and G. R. Marshall, *Ann. Rep. Med. Chem.*, **15**, 267 (1980).
14. W. C. Herndon and S. H. Bertz, *J. Comput. Chem.*, **8**, 367 (1987).
15. G. Rum and W. C. Herndon, *J. Am. Chem. Soc.*, **113**, 9055 (1991).
16. M. L. Connolly, *Science*, **221**, 709 (1983).
17. P. E. Bowen-Jenkins, D. L. Cooper, and W. G. Richards, *J. Phys. Chem.*, **89**, 2195 (1985).
18. A. M. Richards and J. R. Rabinowitz, *Int. J. Quant. Chem.*, **31**, 309 (1987).
19. P. E. Bowen-Jenkins and W. G. Richards, *Int. J. Quant. Chem.*, **30**, 763 (1986).
20. E. E. Hodgkin and W. G. Richards, *Int. J. Quant. Chem. Quant. Biol. Symp.*, **14**, 105 (1987).
21. R. Carbó, L. Leyda, and M. Arnau, *Int. J. Quant. Chem.*, **27**, 1185 (1980).
22. D. L. Cooper and N. L. Allan, *J. Computer-Aided Mol. Design*, **3**, 253 (1989).
23. A. J. C. Wilson, Ed., *International Tables for Crystallography*, *Vol. C*, Kluwer, Dordrecht, 1992.
24. G. M. Sheldrick, *SHELX(76), Program for Crystal Structure Determination*, University of Cambridge, Cambridge, UK, 1976.
25. J. A. Nelder and R. Mead, *Comput. J.*, **7**, 308 (1965).
26. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
27. J. Badger, I. Minor, M. J. Kremer, M. A. Oliveira, T. J. Smith, J. P. Griffith, D. M. A. Guerin, S. Krishaswami, M. Luo, M. G. Rossmann, M. A. McKinlay, G. D. Diana, F. J. Dutko, M. Fancher, R. R. Rueckert, and B. A. Heinz, *Proc. Natl. Acad. Sci. USA*, **85**, 3304 (1988). PDB code for this compound is 2R04.
28. J. F. Davies, T. J. Delcamp, N. J. Prendergast, K. A. Ashford, J. H. Freisheim, and J. Kraut, *Biochemistry*, **27**, 3664 (1988). The PDB code is 1DHF.
29. J. T. Bolin, D. F. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut, *J. Biol. Chem.*, **257**, 13650 (1982). The PDB code is 4DFR.
30. H. Brandstetter, D. Turk, H. W. Hoeffken, D. Grosse, J. Stürzebecher, P. D. Martin, B. F. P . Edwards, and W. Bode, *J. Mol. Biol.*, **226**, 1085 (1992). PDB codes for MQPA, NAPAP, and 4-TAPAP are 1ETR, 1ETS, and 1ETT, respectively.